Computational Statistics and Data Analysis 90 (2015) 15–35

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Grouped variable importance with random forests and application to multiple functional data analysis

Baptiste Gregorutti [a,b,*], Bertrand Michel [b], Philippe Saint-Pierre [b]

[a] Safety Line, 15 rue Jean-Baptiste Berlier, 75013 Paris, France

[b] Laboratoire de Statistique Théorique et Appliquée, Sorbonne Universités, UPMC Univ Paris 06, F-75005 Paris, France

ARTICLE INFO

Article history:
Received 19 November 2014
Received in revised form 7 April 2015
Accepted 8 April 2015
Available online 17 April 2015

Keywords:
Random forests
Functional data analysis
Group permutation importance measure
Group variable selection

ABSTRACT

The selection of grouped variables using the random forest algorithm is considered. First a new importance measure adapted for groups of variables is proposed. Theoretical insights into this criterion are given for additive regression models. Second, an original method for selecting functional variables based on the grouped variable importance measure is developed. Using a wavelet basis, it is proposed to regroup all of the wavelet coefficients for a given functional variable and use a wrapper selection algorithm with these groups. Various other groupings which take advantage of the frequency and time localization of the wavelet basis are proposed. An extensive simulation study is performed to illustrate the use of the grouped importance measure in this context. The method is applied to a real life problem coming from aviation safety.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

In the high dimensional setting, identification of the most relevant variables has been the subject of much research during the last two decades (Guyon and Elisseeff, 2003). For linear regression, the lasso method (Tibshirani, 1996) is widely used. Many variable selection procedures have also been proposed for nonlinear methods. In the context of random forests (Breiman, 2001), it has been shown that the permutation importance measure is an efficient tool for selecting variables (Díaz-Uriarte and Alvarez de Andrés, 2006; Genuer et al., 2010; Gregorutti et al., 2014).

In many situations such as medical studies and genetics, groups of variables can be clearly identified and it is of interest to select groups of variables rather than to select them individually (He and Yu, 2010). Indeed, interpretation of the model may be improved along with the prediction accuracy by grouping the variables according to *a priori* knowledge about the data. Furthermore, grouping variables can be seen as a solution to stabilize variable selection methods. In the linear setting, and more particularly for linear regression, the group lasso has been developed to deal with groups of variables, see for instance Yuan and Lin (2006a). Group variable selection has also been proposed for kernel methods (Zhang et al., 2008) and neural networks (Chakraborty and Pal, 2008). As far as we know, this problem has not been studied for the random forest algorithm introduced by Breiman (2001). In this paper, we adapt the permutation importance measure for groups of variables in order to select groups of variables in the context of random forests.

The first contribution of this paper is a theoretical analysis of the grouped variable importance measure. Generally speaking, the grouped variable importance does not reduce to the sum of the individual importances and may even be

* Corresponding author at: Safety Line, 15 rue Jean-Baptiste Berlier, 75013 Paris, France.
E-mail addresses: baptiste.gregorutti@safety-line.fr (B. Gregorutti), bertrand.michel@upmc.fr (B. Michel), philippe.saint_pierre@upmc.fr (P. Saint-Pierre).

http://dx.doi.org/10.1016/j.csda.2015.04.002
0167-9473/© 2015 Elsevier B.V. All rights reserved.

quite unrelated to it. However, in more specific models such as additive regression ones, we derive exact decompositions of the grouped variable importance measure.

The second contribution of this work is an original method for selecting functional variables based on the grouped variable importance measure. Functional Data Analysis (FDA) is a field in statistics that analyzes data indexed by a continuum. In our case, we consider data providing information about curves varying over time (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Ferraty, 2011). One standard approach in FDA consists in projecting the functional variables onto a finite dimensional space spanned by a functional basis. Classical bases in this context are splines, Fourier, wavelets or Karhunen–Loève expansions, for instance. Most of the papers about regression and classification methods for functional data consider only one functional predictor; references include Cardot et al. (1999, 2003), Rossi et al. (2006) and Cai and Hall (2006) for linear regression methods, Amato et al. (2006) and Araki et al. (2009) for logistic regression methods, Górecki and Smaga (2015) for ANOVA problem, Biau et al. (2005) and Fromont and Tuleau (2006) for k-NN algorithms and Rossi and Villa (2006) and Rossi and Villa (2008) for SVM classification. The multiple FDA problem, where p functional variables are observed, has been less studied. Recently, Matsui and Konishi (2011) and Fan and James (2013) have proposed solutions to the linear regression problem with lasso-like penalties. The logistic regression case has been studied by Matsui (2014). Classification based on several functional variables has also been considered using the CART algorithm (Poggi and Tuleau, 2006) and SVM (Yang et al., 2005; Yoon and Shahabi, 2006).

We propose a new approach for multiple FDA using random forests and the grouped variable importance measure. Indeed, various groups of basis coefficients can be proposed for a given functional decomposition. For instance, one can choose to regroup all coefficients of a given functional variable. In this case, the selection of a group of coefficients corresponds to the selection of a functional variable. Various other groupings are proposed for wavelet decompositions. For a given family of groups, we adapt the recursive feature elimination algorithm (Guyon et al., 2002) which is particularly efficient when predictors are strongly correlated (Gregorutti et al., 2014). In the context of random forests, this backward-like selection algorithm is guided by the grouped variable importance. Note that by regrouping the coefficients, the computational cost of the algorithm is drastically reduced compared to a backward strategy that would eliminate only one coefficient at each step.

An extensive simulation study illustrates the application of the grouped importance measure for FDA. The method is then applied to a real life problem coming from aviation safety. The aim of this study is to explain and predict landing distances. We select the most relevant flight parameters regarding the risk of long landings, which is a major issue for airlines.

The group permutation importance measure is introduced in Section 2. Section 3 deals with multiple FDA using random forests and the grouped variable importance measure. The application to flight data analysis is presented in Section 4. Note that additional experiments about the grouped variable importance are given in Appendix B. In order to speed up the algorithm, the dimension of the data can be reduced in a preprocessing step. In Appendix C, we propose a modified version of a well-known shrinkage method (Donoho and Johnstone, 1994) that simultaneously shrinks to zero the coefficients of the observed curves of a functional variable.

2. The grouped variable importance measure

Let Y be a random variable in $\mathbb{R}$ and $\mathbf{X}^\top = (X_1, \ldots, X_p)$ a random vector in $\mathbb{R}^p$. We denote by $f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ the regression function. Let $\mathrm{Var}(\mathbf{X})$ and $\mathrm{Cov}(\mathbf{X})$ denote the variance and variance–covariance matrices of $\mathbf{X}$.

The permutation importance introduced by Breiman (2001) measures the accuracy of each variable X_j for predicting Y. It is based on the elementary property that the quadratic risk $\mathbb{E}\left[(Y - f(\mathbf{X}))^2\right]$ is the minimum error for predicting Y knowing $\mathbf{X}$. The formal definition of the variable importance measure of X_j is:

$$\mathcal{I}(X_j) := \mathbb{E}\left[\left(Y - f(\mathbf{X}_{(j)})\right)^2\right] - \mathbb{E}\left[(Y - f(\mathbf{X}))^2\right], \tag{1}$$

where $\mathbf{X}_{(j)} = (X_1, \ldots, X_j', \ldots, X_p)^\top$ is a random vector such that X_j' is an independent replicate of X_j which is also independent of Y and of all other predictors. This criterion evaluates the increase of the prediction error after breaking the link between the variable X_j and the outcome Y (see Zhu et al., 2012 for instance).

In this paper, we extend the permutation importance to groups of variables. Let $J = (j_1, \ldots, j_k)$ be a k-tuple of increasing indices in $\{1, \ldots, p\}$, with $k \leq p$. We define the permutation importance of the sub-vector $\mathbf{X}_J = (X_{j_1}, X_{j_2}, \ldots, X_{j_k})^\top$ of predictors by

$$\mathcal{I}(\mathbf{X}_J) := \mathbb{E}\left[\left(Y - f(\mathbf{X}_{(J)})\right)^2\right] - \mathbb{E}\left[(Y - f(\mathbf{X}))^2\right],$$

where $\mathbf{X}_{(J)} = (X_1, \ldots, X_{j_1}', X_{j_1+1}, \ldots, X_{j_2}', X_{j_2+1}, \ldots, X_{j_\ell}', X_{j_\ell+1}, \ldots, X_p)^\top$ is a random vector such that $\mathbf{X}_J' = (X_{j_1}', X_{j_2}', \ldots, X_{j_k}')^\top$ is an independent replicate of $\mathbf{X}_J$, which is also independent of Y and of all other predictors. We call this quantity the grouped variable importance since it only depends on which variables appear in $\mathbf{X}_J$. By abuse of notation and ignoring the ranking, we may also refer to $\mathbf{X}_J$ as a group of variables.

Download English Version:

https://daneshyari.com/en/article/416285

Download Persian Version:

https://daneshyari.com/article/416285

Daneshyari.com